OVERSAMPLING TECHNIQUES

SMOTE (Synthetic Minority Over-sampling Technique)

- SMOTE is one of the most widely used and well-known oversampling techniques for imbalanced datasets.

- It works by generating new synthetic minority class samples by interpolating between existing minority class instances and their nearest neighbors.

- The key idea is to create new samples that fill in the feature space between existing minority class instances, making the class boundaries more distinct.

- SMOTE has been shown to be effective in many applications, but it can sometimes create synthetic samples in ill-fitted locations if the minority class is scattered or has a complex distribution.

ADASYN (Adaptive Synthetic Sampling Approach)

- ADASYN is an extension of SMOTE that aims to address some of its limitations.

- ADASYN adaptively adjusts the number of synthetic samples generated for each minority class instance based on its density distribution.

- It generates more synthetic samples for minority class instances that are harder to learn, i.e., those with fewer neighboring instances of the same class.

- This focus on generating more samples for outlier or scattered minority class instances can improve performance compared to SMOTE in some cases.

- ADASYN has been shown to be particularly effective when the minority class has a complex or multi-modal distribution.

MAHAKIL

- MAHAKIL is a more recent diversity-based oversampling approach that uses Mahalanobis distance to divide minority class instances into two groups.

- It then generates new synthetic samples by performing "crossover" interpolation between pairs of instances from the two groups, creating more diverse synthetic samples.

- The key idea behind MAHAKIL is to improve upon SMOTE and ADASYN by generating more representative and diverse synthetic samples, especially when the minority class is small compared to the feature dimensionality.

- MAHAKIL has been shown to outperform SMOTE and ADASYN in some scenarios, particularly when the minority class has a complex or scattered distribution.

Current State-of-the-Art

- The field of oversampling techniques for imbalanced datasets is continuously evolving, with more advanced methods being developed.

- Some recent state-of-the-art techniques include Borderline-SMOTE and MWMOTE (Majority Weighted Minority Oversampling Technique).

- Borderline-SMOTE focuses on identifying and selectively oversampling the more important or "borderline" minority class instances to create better synthetic samples.

- MWMOTE combines oversampling and undersampling techniques, using a weighted approach to generate synthetic minority class samples and remove majority class instances.

- The current state-of-the-art often involves ensemble methods that combine multiple oversampling and undersampling techniques to achieve the best performance on imbalanced datasets.

In summary, SMOTE, ADASYN, and MAHAKIL represent different approaches to generating synthetic minority class samples, with ADASYN and MAHAKIL aiming to address some of the limitations of SMOTE. The field continues to evolve, with more advanced techniques being developed to handle increasingly complex imbalanced datasets.